A computer-assisted approach to the analysis of metaphor variation across genres

Elena Semino, Andrew Hardie, Veronika Koller and Paul Rayson
Departments of Linguistics and Computing, Lancaster University

Introduction

The importance of metaphor in scientific discourse is widely recognized, and some earlier studies have considered variation in metaphor use in different genres dealing with scientific issues (e.g. Nerlich and Halliday 2007, Semino 2008: 140ff.). More specifically, Skorczynska and Deignan (2006) have used a corpus-based approach to investigate the differences in the metaphors used in academic vs. popular business discourse. However, Skorczynska and Deignan's study, like many others in recent metaphor research, relies on the concordancing of preselected search strings, so that the analysts can, by and large, only find further instances of expressions they have already identified as worthy of study.

Our project exploits for metaphor analysis a software package (Wmatrix) that was originally developed by Paul Rayson for the semantic analysis of texts (Rayson 2008). The USAS tool within Wmatrix automatically allocates each word or multi-word expression in a data set to a semantic field, and allows the analyst to compare the frequency of semantic fields in a particular data set with their frequency in a relevant reference corpus. The semantic fields exploited by USAS can be seen as roughly corresponding to the domains of metaphor theory. In our approach, we are able to filter all possible candidate semantic fields proposed by USAS to assist in finding possible 'source' (e.g. WAR) and 'target' (e.g. MEDICAL TREATMENT) domains, and we can then go on to consider the potential metaphoricity of the expressions included under each possible source domain. This method thus enables us to identify open-ended sets of metaphoric expressions, which are not limited to pre-determined search strings. We have already successfully employed this approach in two pilot studies involving different genres and discourses (see Hardie *et al.* 2007, Koller *et al.* 2008).

In this paper we present some of the results of a small-scale corpus-based study of variation in metaphor use in two genres that can both be subsumed under the broad label of 'scientific discourse'. The first data-set (approximately 19,000 words) consists of six articles included in a special issue of the scientific journal *Nature Immunology*, published in 2005. The special issue was devoted to the topic of 'regulatory T cells', a type of lymphocytes which have the crucial function of preventing the body's immune system from reacting against the body's own healthy cells, while reacting against harmful external agents. The use of metaphor in the special issue is considered in Semino (2008: 157-63). The second data-set (approximately 14,000 words) is a collection of 15 articles that appeared in the *New Scientist* between 2005 and 2009, and that were identified via a search for the string 'regulatory T-cells' in the magazine's online archive. The genres represented by the two data-sets deal with the same broad topic area, but differ in terms of both authors and audiences: the articles from *Nature Immunology* are written by experts for an audience of experts, while the articles from the *New Scientist* are written by science reporters for an educated but more general audience.

The analysis presented in this paper shows how our approach enables us to identify differences between the two datasets, particularly in terms of (a) the frequency with which particular source domains are exploited, and (b) the similarities and differences in the ways in which the same source domains are realised in the two sub-corpora.

Method

Our approach began by automatically applying two levels of linguistic annotation to the texts in the *New Scientist* and *Nature Immunology* datasets. First, we assigned part-of-speech tags using the CLAWS tagger (Garside and Smith 1997) and then exploited those distinctions in order to assign semantic field tags using the USAS system (Rayson *et al.* 2004). The USAS categories group words and multiword expressions into their general field of meaning using a taxonomy

derived from the labels in the Longman Lexicon (McArthur 1981). It should be noted that the USAS domains were not designed with metaphor analysis in mind, but in our method they are used to assist the metaphor researcher to identify groups of significant words and multiwords for further analysis via concordancing.

The second step in the analysis was to apply keyness statistics in order to determine which semantic categories have unusually high or low frequencies in the two datasets. This approach has been pioneered at the level of word frequencies in WordSmith keywords (Scott 2008). However, we used the Wmatrix software, which calculates a similar statistic for tag frequencies. Wmatrix computes the relative difference in the frequency of each semantic tag in a given dataset against a reference corpus. In this paper, we compared the two datasets directly, thus using each dataset as a reference for the other to provide expected frequencies. The statistic used is Log Likelihood (LL), which is the same one used to extract keywords in WordSmith. The larger the LL value, the more significant the difference in relative frequencies between the two datasets. LL values over 6.63 are significant at a probability value of 0.01 (1 degree of freedom), indicating that the difference observed is highly unlikely to be due to random chance. This process is explained in further detail in Rayson (2008).

We thus extracted a list of the semantic categories most significant (and with LL $\geq$ 6.63) in each dataset relative to the other. Amongst these, we identified those categories likely to operate as metaphoric source domains (see Table 1, which shows the code and definition of each semantic category, together with its log-likelihood value).

Possible metaphoric source domains among overused semantic domains in New Scientist vs. Nature Immunology data	*Possible metaphoric source domains among overused semantic domains in Nature Immunology vs. New Scientist data*
E3- violent/angry (53.90)	**S7.4-** not allowed (62.64)
Q2.1 speech: communicative (35.46)	**Q1.1** linguistic action (36.0)
A1.7- no constraints (31.78)	**Q3** language, speech and grammar (34.72)
M1 moving, coming and going (24.13)	**S7.4+** allowed (20.07)
M4 sailing, swimming (16.05)	**X5.2+** interested/excited/energetic (15.72)
M2 putting, pulling, pushing, transporting (12.07)	**A1.7+** constraint (13.39)
W4 weather (12.04)	**S1.2.2+** greedy (12.79)
O4.5 texture (12.04)	**S1.2.6-** foolish (11.08)
K1 entertainment generally (10.27)	
A1.5.1- unused (8.93)	
A10- closed; hiding, hidden (7.66)	
G3 warfare, defence and the army; weapons (7.31)	

Table 1: Semantic domains overused in each dataset relative to the other (log likelihood values in brackets)

Already at this very broad-focussed level of analysis, we can discern possible differences in metaphor use between the two datasets, as well as similarities. For example, the very fact that a wider range of potential source domains can be identified for the *New Scientist* data suggests that a greater *variety* of metaphor may be a feature of the texts written for a general rather than technical audience. However, this hypothesis needs to be confirmed in the course of a more detailed analysis.

In some cases, the prominence of a semantic category was traceable to the prominence of just a single lexical item within that category. For example, the category A1.5.1- 'Unused' is exemplified in the *New Scientist* data solely by the word *dormant*. However, in most cases the domains identified in Table 1 represented multiple words. Therefore, the next step in the analysis of the metaphors in use across the two datasets is to examine the underlying lists of word types for each potential source domain. Taking the word types into account allows for a more finely-grained analysis of the forms of metaphor being employed. Furthermore, the frequencies of different word types within a source domain tells us something about the reasons for that domain's keyness: is it the cumulative effect of a wide range of terms related to that domain, or

is the high frequency driven largely by a much smaller number of word types? As we will exemplify below, this is often an important issue for an analysis of the sort we undertake here. Crucially, the word lists for a given domain in *both* datasets were examined in contrast, the dataset where it is relatively less frequent *as well as* the dataset where it is relatively prominent. Often, it is an examination of the contrasts between the word lists that brings into sharpest relief what is unusual or notable about the metaphoric usage of that particular domain in the dataset where it is key.

The individual key semantic domains, or tags, can be broken down into lemmas, which again feature various word types or forms. Finally, each type shows a specific number of occurrences, or tokens. These four embedded layers allow for different ratios to be calculated in order to arrive at different information about how a domain is realised at the surface level of language. Given our overall interest in this paper in the variation of metaphor source domains across genres, we here restrict ourselves to the tag-lemma ratio,[1] which indicates how varied the lexical realisation of a semantic domain is, and the lemma-token ratio. Note that the lemma-token ratio is here not used to measure lexical variety in the corpus as a whole but is arrived at by dividing the number of different lemmas realising a domain by the number of occurrences of those lemmas. Thus, it here measures frequency in the usage of particular lemmas.

Let us take, as a relatively straightforward example of how our analysis operated, the case of the semantic domains relating to personal qualities, namely S1.2.6- and S1.2.2+ ('foolish' and 'greedy' respectively). These are both key in the *Nature Immunology* data. Taken at face value, their presence as key domains would suggest a greater degree of metaphoric personification in the *Nature Immunology* data than in the *New Scientist*. However, refining the analysis by looking at word lists and concordances for these domains results in a much subtler picture. The only two 'greedy' words are *avidity* and *avidities* – a set technical metaphor conceptualising biochemical attraction in terms of greed. In fact, most instances of these terms occur in conjunction with some sort of quantifying modifier, frequently a form of *high* or a close synonym, suggesting that this metaphoric expression is also strongly contextually and/or functionally constrained:

```
of relatively high T cell receptor  avidity  for their selecting molecules witho
5-CD4+ T cells showed an increased  avidity  of Treg cell TCRs for self peptide
onse to self-reactive TCRs with an  avidity  range for self peptide ? MHC somewh
that a TCR signal within a certain  avidity  range is required but not sufficien
```

The 'foolish' words are *naive* in *Nature Immunology*, and *duped* and *fooling* in *New Scientist*. *Naive* (23 instances) is always used to modify either *t cells* or *mice*; again, these are not actually instances of active personification, but rather fixed metaphoric expressions that act simply as technical labels for a particular sort of T cell and model animal respectively. The contrasting words in *New Scientist* occur as follows:

```
ber 2005 by John Pickrell , Dublin  Fooling  the human immune system into accept
lls to fight disease , Platt says .  Duped  cells may prevent transplant rejecti
. " One way to produce an armada of  duped  cells for protecting grafted tissues
ethod in mice . The process - which  dupes  the cells the human body uses to reg
```

These concordance lines clearly show the conceptualisation of cells as sentient in order to communicate the mode of action of a particular medical treatment. So we see that, on the one hand, in the *Nature Immunology* data, the personification metaphor is used in fixed, non-productive, labelling terminology (and used quite frequently); but on the other hand, in the *New Scientist* data, it is used less frequently, but with a slightly greater variety of instantiating expressions, and with an actual, productive conceptual mapping.

[1] Obviously, the numerical value for all tags is 1.

This basic analysis has illustrated the means by which key domain data can be analysed to draw distinctions between the kinds of metaphoric resources called on within each genre. Of course, looking at this relatively isolated fragment of key domain data does not provide anything like a proper overview of the datasets as a whole, although it does suggest some tantalising hints. However, it serves to demonstrate the approach we have taken to the data, and to highlight in particular the importance of several points: (a) the interpretation of the key domains is dependent on investigation of the word forms within each domain and the context those forms occur in; (b) inspecting a key domain in *both* datasets is necessary to draw out the distinctions between them in a meaningful way; and (c) the distribution of tokens identified as metaphoric across word types can inform a contrast of their metaphoric usage in terms of fixedness and productivity. Having illustrated the method and these points, we move on to a consideration of two complexes of interacting key domains indicative of particularly significant metaphoric source domains.

Analysis

Source domains: WAR, VIOLENCE/ANGER AND CONSTRAINT

When comparing the two sub-corpora of scientific and popular science articles against each other, a number of semantic domains are indicated as being overused in one sub-corpus or the other with statistically significant frequency. Of these key domains, we will in the following focus on two sets of conceptually related domains, the first of which comprises domains related to WAR, VIOLENCE/ANGER and CONSTRAINT. These are grouped according to conceptual links in Table 2, along with their log-likelihood value. For the sake of comparison, all relevant domains are also listed for the corpus in which they are not key, i.e. significantly underused. Non-keyness is indicated by dashes.

semantic domain	**Nature Immunology (NI)**	**New Scientist (NS)**
E3- violent/angry	--	53.90
G3 warfare, defence and the army; weapons	--	7.31
A1.7+ constraint	13.39	--
A1.7- no constraint	--	31.78
S7.4+ allowed	20.07	--
S7.4- not allowed	62.64	--

Table 2: Key semantic domains WAR, VIOLENCE/ANGER and CONSTRAINT in the two sub-corpora

It is worth noting that the relatively high log-likelihood scores for 'violent/angry' and 'no constraint' in the *New Scientist* point to a dominance of these concepts that is also reflected when comparing the popular science articles against the written sampler of the British National Corpus (BNC); here, 'no constraint' scores as high as 165.45, while 'violent/angry' still records a log-likelihood of 15.67. Conversely, the domain that is most significantly overused in the *Nature Immunology* texts, 'not allowed', comes up at 577.67 when one compares the sub-corpus against the BNC written sampler. (Incidentally, while 'not allowed' is not a key domain in the New Scientist when compared to *Nature Immunology*, the domain is significantly overused in the former when compared against the BNC written sampler, albeit at the much lower log-likelihood score of 21.49.)

Overall, the *New Scientist* records 64 key semantic domains when compared against *Nature Immunology*, while the latter only features 35 overused domains when compared against the *New Scientist*. This suggests that the popular science articles are based on a much broader range of semantic concepts. As a result, the *New Scientist* key semantic domains 'violent/angry', 'warfare etc.' and 'no constraint' are not only conceptually related to each other, but also to particular realisations of other key domains. Thus, the domain 'sailing, swimming' (LL 16.05) includes two occurrences of the word 'armada', while the domain 'no obligation or necessity' (LL 19.95) features the phrases 'free of cancer' and 'sets the immune system free to fight cancer

cells'. Also, the semantic domain 'damaging and destroying' falls just short of statistical significance, at 6.47. The pervasiveness of these two related concepts throughout the key domains is another indication of their dominance in the *New Scientist* sub-corpus.

The tag-lemma and lemma-token ratios are presented in Table 3, which again indicates whether and to what extent a specific domain is key in of the sub-corpora.

semantic domain (tag)	**sub-corpus**	**number of lemmas**	**number of occurrences (tokens)**	**tag-lemma ratio**	**lemma-token ratio**
E3- violent/angry	NI (--)	5	7	0.20	0.71
	NS (53.90)	12	44	0.08	0.27
G3 warfare, defence	NI (--)	1	3	1.00	0.33
and the army; weapons	NS (7.31)	6	9	0.17	0.67
A1.7+ constraint	NI (13.39)	8	56	0.13	0.14
	NS (--)	5	11	0.20	0.45
A1.7- no constraint	NI (--)	4	45	0.25	0.09
	NS (31.78)	5	74	0.20	0.07
S7.4+ allowed	NI (20.07)	3	61	0.33	0.05
	NS (--)	5	9	0.20	0.56
S7.4- not allowed	NI (62.64)	3	152	0.33	0.02
	NS (--)	3	17	0.33	0.18

Table 3: Tags, lemmas and tokens in the two sub-corpora (domains WAR, VIOLENCE/ANGER and CONSTRAINT)

What we can observe in Table 3 is first of all a certain frequency effect, where one lemma shows a high number of occurrences, thus making the domain key. In particular, this can be seen for 'no constraint', 'allowed' and 'not allowed', which show the lowest lemma-token ratios for key domains. However, this finding is counterbalanced by variety in the lexicalisation of domains. Table 3 also shows that while the respective lemmas are used less often when a domain is key, the domain itself tends to be realised by a higher number of lemmas. Interestingly, the two cases where a key domain is not more, or even less, lexically varied than its non-key counterpart in the other sub-corpus are 'allowed' and 'not allowed', which are key in *Nature Immunology*. All *New Scientist* key domains are realised by a larger number of lemmas than their non-key equivalents in *Nature Immunology*. This suggests that the popular science articles could be more lexically varied overall, which would mirror the greater semantic variety this sub-corpus shows: As noted above, the *New Scientist* texts feature about a third more key semantic domains than the *Nature Immunology* articles. Indeed, the former sub-corpus shows a type-token ratio of 0.193 overall, while the scientific corpus records only 0.131.[2] This higher lexical variety is balanced by lower frequency of use though; the only case where lexical items realising a key domain are used with higher frequency is that of 'warfare etc.', a central concept in the *New Scientist* sub-corpus. The lemma-token ratio we have established for the WAR source domain in the popular science texts replicates the one that Downing and Mujic (2009: 70) have ascertained for the use of that domain in another popular science journal, *Scientific American*.

Qualitatively speaking, the lemmas realising the domains of WAR, VIOLENCE/ANGER and CONSTRAINT in *Nature Immunology* are indeed all used metaphorically in every single occurrence. This is not always the case for the lemmas realising the same domains in the *New Scientist* corpus; for example, lemmas tagged as 'violent/angry' and 'allowed' include a small number of misidentifications due to polysemy and formulaic usage ('substances such as dust mites and cat dander', 'Let us hope this tragedy tells us more'). Other items were correctly identified but used literally (e.g. '[people] have to be attached to kidney machines' realising the 'constraint' domain). More interestingly, a few items in the central 'violent/angry' domain were

[2] The ratio here is arrived at by dividing the number of types (word forms) and multi-word units by the number of tokens, excluding abbreviations, numbers and proper names. Wmatrix does not currently show the number of different lemmas in a corpus.

misidentified by being given a first tag for metaphoric meaning. Thus, *venom(s)* and *ferment* were classified as lemmas realising the domain of VIOLENCE/ANGER when they are in fact used literally, as in 'purified bee venom' or 'to ferment the pepperoni'. This tag order indicates the degree to which the respective items have become conventionalised as metaphoric expressions (see Hardie *et al.* 2007, Koller *et al.* 2008).[3]

The exclusively metaphoric usage of all lemmas in the relevant *Nature Immunology* domains suggests that the scientific texts use a smaller number of technical metaphoric expressions, such as 'aggressive lymphoproliferation' or 'suppressive function of natural Treg cells'. By contrast, the popular science genre tends to adopt one of the high-frequency lemmas from the scientific articles and expand on it by introducing further, less precise and more colloquial, realisations of the same domain. For example, the technical term *aggressive* is used alongside the much more frequent *attack* and is also expanded by lemmas like *berserk*, *hits* and *knocked off*. Similarly, the metaphoric technical term *invaders* is used together with *bombarded*, weapons and *front-line troops*. Of the three examples given below, the first two are from our *New Scientist* sub-corpus, while the third is taken from Downing and Mujic's study (2009) of metaphor in the abstracts of articles on immunology published in *Scientific American*, showing that the WAR metaphor may be routinely extended in popular science writing.

(1) Recent breakthroughs in our understanding of how the immune system's front-line troops recognise the enemy are shedding light on how contact with harmless microbes might play a big role in teaching the immune system when to attack and when to retreat. (*New Scientist* sub-corpus)
(2) [W]e know that less aggressive tumours contain large numbers of killer T-cells and memory T-cells , which are both powerful weapons of the immune system, suggesting that it is mounting an effective response. (*New Scientist* sub-corpus)
(3) Normally, the immune system is able to distinguish friend from foe, ignoring the body's own components and attacking foreign invaders. Unfortunately, the immunological weapons can, like friendly fire, sometimes turn against the self. (Downing and Mujic 2009: 76)

In our sub-corpus, the items underlined above are all tagged as 'violent/angry' and 'warfare etc.', respectively, whereas others not listed here − *out of control* and *run amok* − are tagged as 'no constraint', showing the conceptual links between the domains.

Source domain: COMMUNICATION

The lists of key semantic domains we obtained by comparing the two sub-corpora against each other include three domains that are broadly related to language and communication. More specifically, Q.21 'speech: communicative' is the eleventh most overused domain in the *New Scientist* sub-corpus when compared with the *Nature Immunology* sub-corpus, while Q1.1 'linguistic actions, states and processes; communication' and Q3 'language, speech and grammar' are, respectively, the seventh and eighth most overused domains in the *Nature Immunology* sub-corpus when compared with the *New Scientist* sub-corpus. As shown in Table 4, the relevant LL values range between 34 and 36.

The dominance of these domains in the two sub-corpora is confirmed by a comparison of each sub-corpus against the written sampler of the BNC: When comparing the *New Scientist* data against the BNC, Q2.1 'speech: communicative' is overused (LL 12.42); when comparing *Nature Immunology* against the BNC, both Q1.1 'linguistic actions, states and processes' and Q3 'language, speech and grammar' are overused (respectively, LL 54.86 and 25.18).

[3] If we amend the quantitative results to only include lemmas correctly identified as metaphoric expressions, the only ratios to change are the tag-lemma ratios for the non-key domains 'allowed' and 'not allowed' in the *New Scientist*. The former would then be the same as for *Nature Immunology* (0.33) and the latter would at 0.50 be lower than the ratio for *Nature Immunology*. In fact, this aligns these domains more closely with the overall trend that key domains are more lexically varied.

semantic domain	Nature Immunology (NI)	New Scientist (NS)
Q2.1 speech: communicative	--	35.46
Q1.1 linguistic actions, states and processes; communication	36.00	--
Q3 language, speech and grammar	34.72	--

Table 4: Key semantic domains SPEECH: COMMUNICATIVE, LINGUISTIC ACTIONS and LANGUAGE, SPEECH AND GRAMMAR in the two sub-corpora

Table 5 provides more detail on the numbers of lemmas and tokens included under each domain. The main pattern that is revealed by the table concerns the relative lexical variety of the relevant domains in the two sub-corpora: Regardless of keyness, the *New Scientist* sub-corpus has a higher lemma-token ratio for all three domains than the *Nature Immunology* sub-corpus. This indicates that the communication-related vocabulary used in the *New Scientist* is more varied than in the case of *Nature Immunology*. This observation is consistent with some of the patterns we noted in earlier sections, but further investigation is needed in order to discover to what extent this involves metaphoric expressions.

semantic domain (tag)	sub-corpus	number of lemmas	number of occurrences (tokens)	tag-lemma ratio	lemma-token ratio
Q2.1 speech: communicative	NI (--)	8	106	0.13	0.08
	NS (35.46)	14	133	0.07	0.11
Q1.1 linguistic actions, states and processes; communication	NI (36.00)	14	128	0.07	0.11
	NS (--)	9	22	0.11	0.41
Q3 language, speech and grammar	NI (34.72)	5	75	0.20	0.07
	NS (--)	5	7	0.20	0.71

Table 5: Tags, lemmas and tokens in the two sub-corpora (domains SPEECH: COMMUNICATIVE, LINGUISTIC ACTIONS and LANGUAGE, SPEECH AND GRAMMAR)

We regarded all three communication-related domains as potential source domains for two reasons. First, their dominance is not straightforwardly justified by the kinds of texts included in the two sub-corpora and their subject matter. Second, previous studies have shown that some chemical processes happening at the level of our body's cells are conventionally expressed in terms of communication (see Semino 2008: 157ff. for an analysis of the *Nature Immunology* data). A more detailed investigation of one of the three domains confirmed our initial expectation, but also revealed some further patterns and complications.

A comparison of the word lists for Q2.1 'speech: communicative' in the two sub-corpora shows that its keyness in the *New Scientist* data is due to a single lemma: *says*. This verb occurs 80 times in the *New Scientist* sub-corpus, where it is used in the reporting of the (spoken or written) words of authoritative sources. Such uses are, of course, non-metaphoric. There are no occurrences of *says* in the *New Scientist* sub-corpus, where different conventions are used to quote or paraphrase from other texts. If we exclude *says* in the *New Scientist* data, however, the vast majority of tokens that were included under the Q2.1 'speech: communicative' domain realise only two types and a single lemma, namely, the noun *response* in its singular and plural forms (i.e. *response* and *responses*). This noun is best described as a highly conventional metaphoric expression that functions as a technical term, as shown by the following extract from the concordance of *responses* in the *Nature Immunology* data:

```
rimental conflicts between immune responses to self versus non-self. A fundam
y the active regulation of immune responses through cellular interactions and
ause Treg cells can dampen immune responses , they may hamper effective contro
o the host . In addition , memory responses against particular pathogens such
```

The concordance lines show that *responses* is conventionally used metaphorically to refer to cellular reactions, and particularly to various ways in which the immune system may become active in order to prevent the damage that may be caused by the presence of particular agents within the body. As shown by the extract above, *responses* predominantly co-occurs with *immune* in the *Nature Immunology* data. The same pattern was found, for both the singular and the plural form of the noun, in both sub-corpora. However, two differences between the two sub-corpora are worth mentioning here.

The first difference concerns frequency: As a lemma, *response* occurs more frequently in the *Nature Immunology* data than in the *New Scientist* data, namely 90 times as opposed to 34 (out of a total of, respectively, 106 and 133 tokens in the two sub-corpora for the semantic domain as a whole). If we consider that 80 of the 133 tokens for Q2.1 'speech: communicative' in the *New Scientist* data are instance of *says*, we have to conclude that, as a metaphoric source domain, Q2.1 'speech: communicative' is more dominant in the *Nature Immunology* sub-corpus than in the *New Scientist* sub-corpus. This is primarily due to the technical status of the metaphoric noun *response*.

The second difference concerns the way in which *response* and *responses* are used as metaphoric expressions. In the *Nature Immunology* sub-corpus, the potential relationship between the technical use of 'response' and what we may call the COMMUNICATION source domain is never overtly exploited in the co-text. This underlines the fact that the noun is used as a technical term in *Nature Immunology*: Neither writers nor readers are likely to rely on the source domain in order to use and understand the term as it is used in communication among specialists. The *New Scientist* sub-corpus, however, contains several instances in which the use of other metaphoric expressions in the co-text suggests that the source domain is being deliberately exploited by the writers (see Cameron 2003: 100ff., Steen 2008):

(4) So where a normal person's immune system flares up when stung, producing inflammation and pain, a venom-tolerant beekeeper's extra regulatory T-cells ensure the response is much more muted.
(5) Within a week, their immune system had muzzled this response, only for it to return the following year after the winter break.

In other words, as we have already noted in relation to other domains, the *New Scientist* writers adopt many of the technical metaphors that we have found in the *Nature Immunology* sub-corpus, but often use them in ways that can be described as deliberate and creative, and that suggest a much greater degree of involvement of the source domain.

Conclusion

Overall, the output of the USAS tool in the two-way comparison of our sub-corpora has allowed us to notice the prominence of several domains that may potentially function as metaphoric source domains and that may provide metaphoric technical terms. A more detailed investigation of word lists and concordances has enabled us to verify the extent to which this is the case, and to note some differences between the two sub-corpora that can be explained in terms of differences between different genres. We have shown specifically how the technical metaphors of specialized scientific genres may be adopted to different extents in scientific writing for more general audiences, and may also be 'opened up' (Knudsen 2003) via the deliberate and potentially creative exploitation of the relevant source domains.

References

Cameron, L. (2003). *Metaphor in Educational Discourse*. London: Continuum.

Downing, L.H. and B.K. Mujic (2009). INFECTIOUS DISEASES ARE SLEEPING MONSTERS: Conventional and culturally adapted new metaphors in a corpus of abstracts on immunology. *Ibérica* 17: 61-82.

Garside, R., and Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In: Garside, R., Leech, G. and T. McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman: London, 102-21.

Hardie, A., Koller, V., Rayson, P. and E. Semino (2007). Exploiting a semantic annotation tool for metaphor analysis. In: *Proceedings of the Corpus Linguistics 2007 Conference*. http://www.corpus.bham.ac.uk/corplingproceedings07/paper/49_Paper.pdf.

Knudsen, S. (2003). Scientific metaphors going public. *Journal of Pragmatics* 35(8): 1247-63.

Koller, V., Hardie, A., Rayson, P. and E. Semino (2008) Using a semantic annotation tool for the analysis of metaphor in discourse. *Metaphorik.de* 15: 141-60. http://www.metaphorik.de/15/koller.pdf.

McArthur, T. (1981). *Longman Lexicon of Contemporary English*. London: Longman.

Nerlich, B. and C. Halliday (2007). Avian flu: The creation of expectations in the interplay between science and the media. *Sociology of Health and Illness* 29(1):46-65.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics* 13(4); 519-49.

Rayson, P., Archer, D., Piao, S.L. and T. McEnery (2004). The UCREL semantic analysis system. In: *Proceedings of the workshop on Beyond Named Entity Recognition Semantic Labelling for NLP Tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004),* 25th May 2004, Lisbon, Portugal, 7-12.

Skorczynska, H. and A. Deignan (2006). Readership and purpose in the choice of economics metaphors. *Metaphor and Symbol* 21(2): 87-104.

Scott, M. (2008). Developing WordSmith. In: Scott, M., Pérez-Paredes P. and P. Sánchez-Hernández (eds.) *Software-aided Analysis of Language*, special issue of *International Journal of English Studies entitled Monograph* 8(1): 153-72.

Semino, E. (2008). *Metaphor in Discourse*. Cambridge: Cambridge University Press.

Steen, G. (2008). The paradox of metaphor: Why we need a three-dimensional model of metaphor. *Metaphor and Symbol* 23: 213-41.